EVALUACIÓN DE RESULTADOS ANALÍTICOS MEDIANTE ANÁLISIS MULTIVARIANTE.

Luis Vicente PÉREZ ARRIBAS
Departamento de Química Analítica
Facultad de C. Químicas
Universidad Complutense de Madrid

Madrid 2018

1.- INTRODUCCIÓN

Como sabemos, el proceso analítico por el cual se toma una muestra y se somete a un tratamiento determinado a partir del cual se obtienen unos resultados, no termina con la realización de estas medidas, si no que se requiere realizar una evaluación de los resultados obtenidos con vistas a una adecuada interpretación de éstos.

Sin embargo, ocurre que en la actualidad los químicos disponemos de avanzados y sofisticados medios instrumentales capaces de suministrar grandes cantidades de datos relativos a los analitos presentes en una muestra, así como de sus propiedades fisico-químicas. Así, p.e., un analizador automático de sangre, de los que se utilizan en un laboratorio clínico es capaz de determinar el contenido en suero y plasma de una veintena de sustancias de interés clínico, que junto con los datos hematológicos y los obtenidos mediante los analizadores de orina pueden suponer un total de unos cuarenta datos clínicos por paciente. De igual manera, cada estación de medición medioambiental del Ayuntamiento de Madrid obtiene cada media hora hasta una veintena de datos atmosféricos y de concentraciones de contaminantes. Esto quiere decir que las veinticinco estaciones que existen en Madrid generan cada día de 15000 a 20000 datos relacionados con la calidad del aire que respiramos. Esto sólo son dos ejemplos, pero podríamos seguir planteando otros muchos más, con lo que es fácil comprender la dificultad que se presenta a la hora de manejar y evaluar esta ingente cantidad de información analítica. El manejo de tan abrumadora cantidad de datos puede ser abordada de dos maneras diferentes:

a) Ignorar la mayoría de los datos de que disponemos y evaluar únicamente aquellos que aportan mayor información
b) Considerar el problema analítico desde un punto de vista multidimensional (o multivariante) y utilizar, si no todos, sí la mayor parte de los datos disponibles.

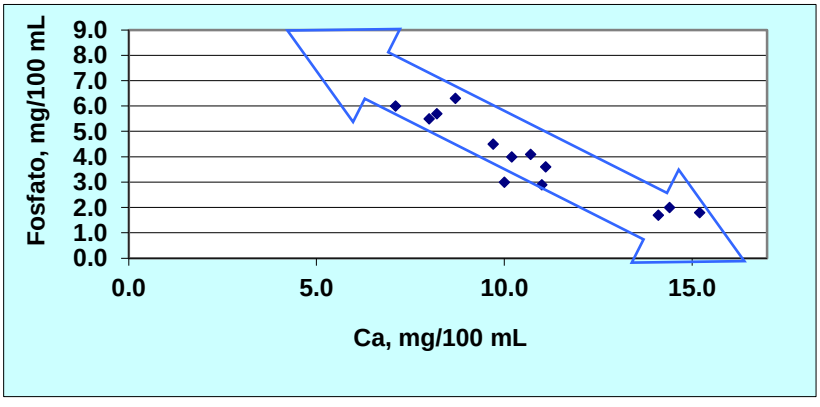
Este artículo tratará de cómo abordar esta multidimensionalidad de los resultados analíticos y la información que de este tratamiento se puede obtener**.** Para ello partiremos de lo que puede ser la definición de **Análisis Multivariante**, que podríamos decir que es el ***conjunto de métodos y herramientas estadísticas utilizadas para evaluar la contribución de un conjunto de factores o variables en los resultados analíticos.***

Para entender este concepto de multimensionalidad, empezaremos por considerar los resultados de un análisis de Ca y PO_4^{3-} en muestras de sangre tomadas a 13 individuos y que se muestran en la siguiente tabla:

Muestra	Ca, mg/100mL	Fosfato, mg/100 mL
1	8,0	5,5
2	8,2	5,7
3	8,7	6,3
4	10,0	3,0
5	10,2	4,0
6	9,7	4,5
7	14,1	1,7
8	15,2	1,8
9	14,4	2,0
10	10,7	4,1
11	11,1	3,6
12	7,1	6,0
13	11,0	2,9
media	10,6	3,9
Mínimo	7,1	1,7
Máximo	15,2	6,3
Desv. Estand.	2,55	1,61

Una evaluación unidimensional de estos resultados analíticos nos permitiría concluir que la concentración de calcio en el suero sanguíneo puede variar entre 7.1 y 15.2 mg/100 mL, o que el fósforo expresado como PO_4^{3-} lo hace entre 1.7 y 6.3 mg/100 mL. Si además, el que evalúa estos resultados analíticos fuera médico, podría concluir, a falta de un posterior reconocimiento clínico, que los individuos 7, 8 y 9 podrían estar sufriendo serias disfunciones del riñón, pues concentraciones altas de calcio en sangre se asocian con enfermedades del riñón como la ***litiasis renal.***

Sin embargo, igual que hemos hecho esta evaluación unidimensional, podríamos haber enfocado el problema desde un punto de vista multidimensional, considerando simultáneamente los resultados obtenidos para el Ca y el PO_4^{3-}. Una forma de afrontar esta evaluación multidimensional podría ser representando las concentraciones obtenidas para el Ca frente a las obtenidas para el P. Esto nos lleva a una gráfica como la de la figura, en la que, como puede apreciarse, los datos aparecen formando tres grupos bastante bien definidos. Además, se puede observar que existe cierta correlación entre los resultados analíticos, ajustándose linealmente a una recta de pendiente negativa que nos da información en cuanto a la forma en que estas dos sustancias se relacionan entre sí. Como vemos, la tendencia general es que cuando la concentración de uno de ellos aumenta, la del otro disminuye y viceversa.



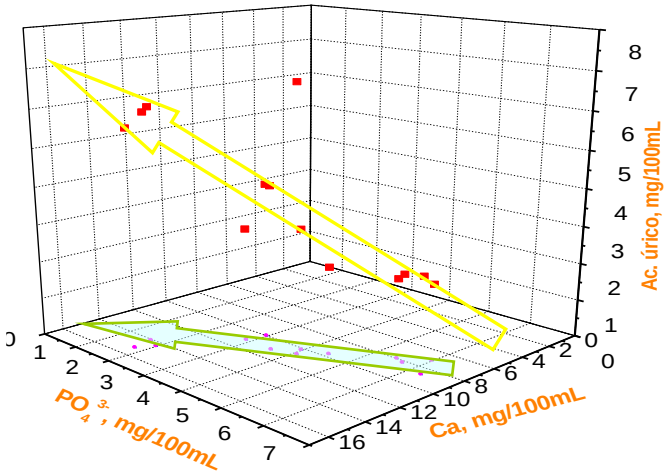
Como vemos, cuando hemos tratado los datos analíticos de forma conjunta, hemos obtenido una información adicional a la que podíamos haber obtenido mediante una evaluación de los resultados analito a analito. Esta evaluación conjunta de los resultados analíticos recibe el nombre de **evaluación multivariante** o más corrientemente **análisis multivariante**, y tiene como uno de sus objetivos principales *el descubrir las relaciones existentes entre las diferentes muestras u objetos analizados, así como entre los analitos incluidos en el análisis*.

Si este análisis multivariante se hace con el único y exclusivo fin de conocer cómo se relacionan entre sí las muestras u objetos analizados, buscando diferencias y similitudes entre ellos, se dice que estamos realizando un **reconocimiento de pautas no supervisado**. Por el contrario, si nosotros conocemos de antemano la relación existente entre las muestras que forman los grupos y las causas que hacen que estos tengan propiedades similares, entonces se dice que el **reconocimiento de pautas es supervisado**. En el caso del ejemplo anterior si nosotros sabemos de antemano que las muestras del grupo central corresponden a individuos sanos, las del grupo inferior a individuos con insuficiencia renal y el superior a otros con problemas de reabsorción de calcio, cualquier otra muestra que nosotros incluyéramos en el estudio junto con éstas, nos permitiría relacionar al individuo con su estado de salud.

2.- ANALISIS DE COMPONENTES PRINCIPALES (PCA)

Igual que hemos relacionado estos dos parámetros relativos a la salud de los individuos, podríamos aumentar el número de estos a ser considerados, pues como ya hemos comentado un análisis de sangre típico incluye entre 15 y 20 parámetros estimadores de la salud de las personas.

Así, p.e, además de Ca y PO_4^{3-} se podría incluir en el estudio el ácido úrico, cuyos valores individuales, por encima de los considerados normales, también están relacionados con determinadas enfermedades renales. Si lo hiciéramos obtendríamos una gráfica tridimensional como la de la figura, en la que, como vemos, nuevamente aparecen agrupamientos de dichas muestras y como en el caso bidimensional, los resultados analíticos muestran cierta correlación entre sí, siguiendo una tendencia determinada.



Igualmente podemos continuar el estudio añadiendo más resultados analíticos correspondientes a otros analitos, como los contenidos en glucosa, bilirrubina, cobre, pH, colesterol, etc. En cualquier caso, dichos resultados experimentales, como paso previo a su evaluación multivariante, deben ser ordenados en forma de matriz, en la cual cada fila corresponde a un **objeto** analizado (por ejemplo un vino, un suelo, una muestra de aire, un individuo, ….) y cada columna a una **magnitud**, **propiedad** o **variable** de interés analítico (por ejemplo, una característica enológica, la concentración de un ion, la concentración de un contaminante, …).

Propiedades → **Objetos** **↓**					
	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	…	$X_{1,k}$
	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	…	$X_{2,k}$
	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$		$X_{3,k}$
	.	.	.		.
	.	.	.		.
	.	.	.		.
	$X_{n,1}$	$X_{n,2}$	$X_{n,3}$	…	$X_{n,k}$

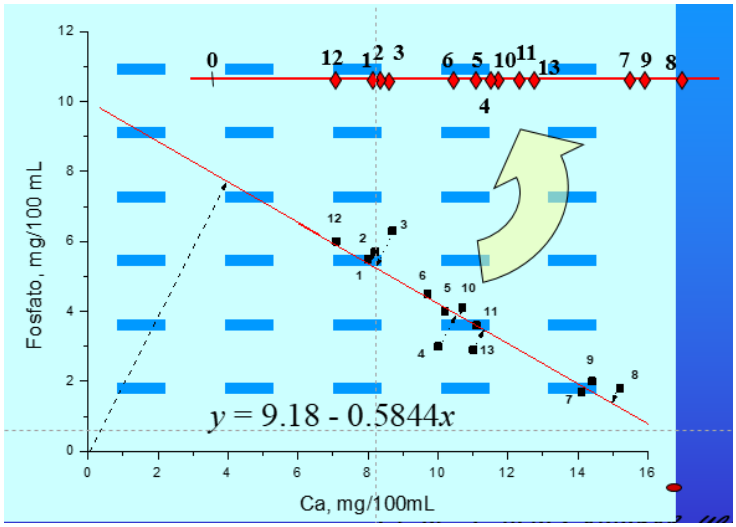
El conjunto de propiedades o variables que caracterizan una muestra u objeto recibe el nombre de **pauta**. Si un grupo de muestras u objetos tienen propiedades similares se dice que ***siguen similares pautas***.

Sin embargo, el mayor problema que se presenta cuando el número de parámetros o variable incluidos es superior a tres, es la imposibilidad de representar los datos en sistemas de coordenadas que nos permitan visualizar las relaciones existentes entre los diferentes objetos analizados. Por consiguiente, se hace necesario buscar la forma de que un sistema multidimensional formado por *k* variables pueda ser reducido a un sistema de coordenadas de cómo máximo tres dimensiones que son las que podemos representar y visualizar.

¿Sería posible esta reducción dimensional sin que por ello se pierda información analítica? Para contestar a esta pregunta, lo mejor es volver al sistema bidimensional del ejemplo primero en el que se intentaba relacionar los contenidos de fosfato y calcio de una serie de muestra de sangre tomadas a 13 individuos. Como ya se vio, estas trece muestras forman tres grupos y se ajustan en mayor o menor grado a una línea recta de ecuación

$$y = 9.18 - 0.5844x$$

Podemos ahora proyectar cada uno de los puntos sobre la recta de regresión, para lo cual trazaremos, procedentes de cada punto y del origen de coordenadas, líneas perpendiculares a la recta de ajuste. Si ahora giramos la recta sobre la que hemos realizado la proyección, hasta una posición horizontal lo que obtenemos es un nuevo sistema de coordenadas unidimensional en el que cada punto del sistema anterior aparece representado sobre una línea recta.



Como vemos, en esta nueva representación los puntos del sistema bidimensional antiguo aparecen de nuevo agrupados de igual manera que antes, y además puesto que la pendiente de la recta de ajuste sobre la que hemos realizado la proyección es negativa, sabemos también que cuando la concentración de uno de los analitos aumenta, la del otro disminuye. En definitiva, lo que hemos hecho es pasar de un sistema de referencia bidimensional a otro monodimensional manteniendo la información fundamental, es decir, las diferencias y similitudes entre las muestras y como se relacionan entre sí los analitos.

A este resultado, al que podemos llegar fácilmente representando los resultados analíticos en papel para gráficos y ayudándonos de una simple calculadora que permita ajustes por mínimos cuadrados, también se puede llegar mediante una serie de sencillos cálculos matemáticos. Para ello sólo hay que multiplicar la coordenada en *x* de cada punto por el coseno del ángulo que forma la recta con la horizontal, y la coordenada en *y* por el seno del mismo ángulo.

$$p_1 = x_1.\cos\theta + y_1.\text{sen}\ \theta$$
$$p_2 = x_2.\cos\theta + y_2.\text{sen}\ \theta$$
$$p_3 = x_3.\cos\theta + y_3.\text{sen}\ \theta$$

Obtenemos así un sistema de ecuaciones que como ocurre con todo sistema puede ser expresado como un producto de matrices:

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = \begin{pmatrix} x_1\ y_1 \\ x_2\ y_2 \\ x_3\ y_3 \end{pmatrix} . \begin{pmatrix} \cos\theta \\ \text{sen}\,\theta \end{pmatrix}$$

o también en notación matricial:

$$\boldsymbol{p = X.l}$$

donde **X** es la matriz que contiene las coordenadas de cada punto y que corresponde con los datos analíticos obtenidos, **p** es un vector matricial que contiene los valores de proyección de cada punto y **l** es un vector formado por el coseno y el seno de θ. Este vector **l** es ortogonal, es decir cumple la condición de que el producto $\mathbf{l}^T.\mathbf{l} = 1$ (uno). El vector **p** recibe el nombre de *vector de "scores"*, palabra de difícil traducción que podría expresarse en español como *vector de valores de proyección ortonormal*; mientras que el vector **l** recibe el nombre de *vector de "loadings"*, traducido a menudo como *vector de carga*, aunque lo correcto desde el punto de vista matemático sería llamarlo *vector de dirección*, pues precisamente representa el vector de dirección de la recta sobre la que se ha realizado la proyección de los puntos.

Siguiendo este procedimiento de cálculo, podemos entonces determinar cuales serían los valores de proyección del ejemplo del análisis de Ca y fosfato en sangre. Para este caso, obtendríamos trece valores de proyección (uno por cada punto o muestra analizada), provenientes de multiplicar las coordenadas de cada punto por el vector de dirección

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ . \\ . \\ p_{13} \end{pmatrix} = \begin{pmatrix} 8.0 & 5.5 \\ 8.2 & 5.7 \\ 8.7 & 6.3 \\ . & . \\ 11.0 & 2.9 \end{pmatrix} . \begin{pmatrix} \cos\theta \\ \text{sen}\,\theta \end{pmatrix}$$

donde θ = artg(m). Puesto que la pendiente de la recta ajustada era -0.5844; θ = -30.30 º.

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ . \\ . \\ p_{13} \end{pmatrix} = \begin{pmatrix} 3.96 \\ 4.03 \\ 4.14 \\ . \\ 7.87 \end{pmatrix} .$$

Si estos valores los representamos sobre un eje, se obtiene la misma representación, con los mismos agrupamientos y distancias al origen que cuando lo hicimos gráficamente.

Puesto que la proyección ortonormal la hemos hecho sobre la recta de regresión a la que mejor se ajustan los datos experimentales, estos valores de proyección reciben el nombre de **valores de proyección del primer Componente Principal** y la recta sobre la que se ha hecho la proyección de los puntos, **1er. Componente Principal**.

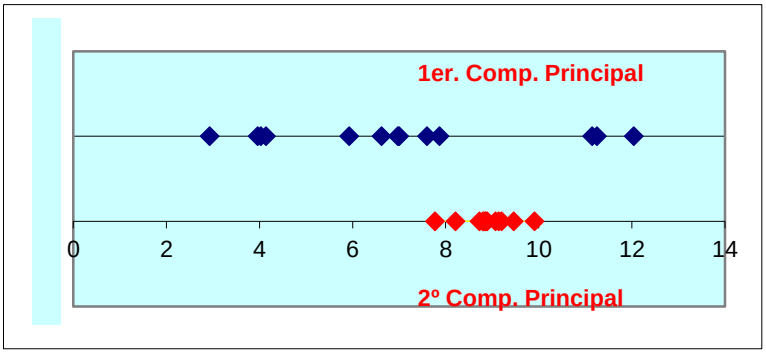
Igual que hemos proyectado los puntos sobre la recta de ajuste de los datos experimentales, podemos proyectar dichos datos sobre cualquier otra recta, por ejemplo sobre una recta que pasando por el origen sea perpendicular a la recta del primer Componente Principal, es decir, sobre su ortonormal.

Como en el caso anterior, la proyección a esta nueva recta puede hacerse gráficamente o también matemáticamente, tomando ahora como ángulo el resultante de sumar 90º al ángulo del 1er. Componente Principal.

$$\boldsymbol{P_2 = X.l_2} = \begin{pmatrix} 8.0 & 5.5 \\ 8.2 & 5.7 \\ 8.7 & 6.3 \\ . & . \\ 11.0 & 2.9 \end{pmatrix} . \begin{pmatrix} \cos(\theta+90) \\ \text{sen}(\theta+90) \end{pmatrix} = \begin{pmatrix} 8.77 \\ 9.04 \\ 9.82 \\ . \\ 7.99 \end{pmatrix} .$$

Obtendríamos así un nuevo vector ***p*** que recibe el nombre de vector de **valores de proyección del 2º Componente Principal**.

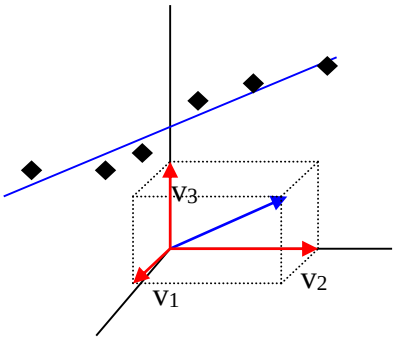
Si representamos y comparamos los valores obtenidos para el 2º Componente Principal con los del 1er. Componente Principal, vemos que ahora los valores de proyección no se agrupan de igual manera y que el intervalo de variabilidad es mucho menor, de aproximadamente 2 unidades, mientras que para el 1er Componente Principal, el intervalo es de unas 9 unidades.



Esto quiere decir que la mayor información sobre los datos analíticos evaluados se encuentra contenida en el 1er. Componente Principal pudiendo prescindirse, hasta cierto punto, de los valores de proyección del 2º Componente Principal. Pero sólo hasta cierto punto, porque el 2º Componente principal aunque aporte poca información, ésta puede ser muy interesante, pues este componente nos da idea del grado de correlación existente entre los datos experimentales. Cuanto más próximos se encuentras entre sí los valores del 2º Componente Principal mayor es la correlación existente entre las muestras. En el caso excepcional que todos los valores del 2º Componente Principal fueran iguales (variabilidad 0), indicaría que los datos experimentales se ajustan perfectamente a una línea recta.

Igual que hemos hecho para una serie bidimensional de datos, podemos hacerlos para una serie tridimensional. Como en el caso anterior, los datos experimentales pueden ajustarse a una línea recta, que al estar en un sistema de tres dimensiones toma la forma:

$$\frac{x-a_1}{v_1}=\frac{y-a_2}{v_2}=\frac{z-a_3}{v_3}$$ (ecuación canóniga de la recta)



donde v_1, v_2 y v_3 son los componentes del vector de dirección de la recta de ajuste de los datos experimentales, por lo que como en el caso bidimensional, una vez que conozcamos estos valores, los valores de las proyecciones sobre el 1er. Componente Principal vendrían dados por el producto de la matriz de datos por este vector de dirección

$$\begin{pmatrix} p_1 \\ p_2 \\ . \\ . \\ p_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ . & . & . \\ . & . & . \\ x_{n,1} & x_{n,2} & x_{n,3} \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

Existen diferentes formas de calcular los valores de v_1, v_2 y v_3. De ellos, quizás el más utilizado por su sencillez sea el procedimiento iterativo de aproximaciones sucesivas denominado **algoritmo NIPALS**, que es el acrónimo inglés **N**onlinear **I**terative **PA**rtial **L**east **S**quares (algoritmo iterativo de mínimos cuadrados parciales no lineales). Este algoritmo es fácil de ejecutar y presenta la ventaja de que simultáneamente al cálculo del vector de dirección se obtienen los valores de proyección sobre el componente principal

Como ya vimos anteriormente, el vector ***p*** se obtiene de multiplicar la matriz de datos ***X*** por un vector de dirección ortogonal, ***l***.

$$\boldsymbol{p = X \cdot l}$$

Como ocurre en todos los métodos iterativos, es necesario dar unos valores arbitrarios de partida, es decir, ***l*** puede ser cualquier vector matricial con tal de que sea ortogonal. Además, para que pueda multiplicar a ***X*** deberá tener tantas filas como columnas tenga la matriz ***X***, o sea, tantas como variables o factores estemos evaluando simultáneamente. Para el caso de tres variables, un vector ortogonal ***l***, de inicio puede ser:

$$l = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad \text{(iniciación)}$$

que es el vector de dirección del eje *x* en un sistema de coordenadas tridimensional y que por definición es ortogonal.

Una vez establecidos los valores de partida, el algoritmo NIPALS comienza calculando un vector de "scores" partiendo de la matriz ***X**** centrada

$$p = X^{*}.l_i \qquad \text{(Paso 1)}$$

y con él un nuevo vector de dirección:

$$l_n^T = p^T.X^{*} \qquad \text{(Paso 2)}$$

Este nuevo vector ***l*** no es ortogonal, por lo que es preciso ortogonalizarlo:

$$l_n = \frac{x_i}{\sqrt{x_1^2 + x_2^2 + x_3^2}} \qquad \text{(Paso 3)}$$

Una vez que ya tenemos ortogonalizado el nuevo vector ***l***, se calcula un nuevo vector de proyecciones, ***p***:

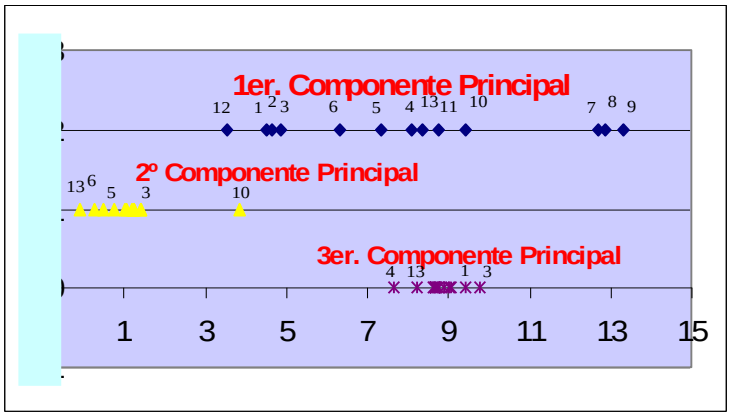
$$p_n = X^{*}.l_n \qquad \text{(Paso 4)}$$

Finalmente, este vector de "scores" se compara con el obtenido en el paso 1. Si la diferencia entre ambos es menor que un valor previamente establecido se toma $\boldsymbol{p}_n$ como el vector de proyecciones del 1$^{er.}$ Componente Principal y $\boldsymbol{l}_n$ como el vector de dirección asociado a este componente. En caso contrario se vuelve al paso 2 y se continúa el proceso hasta cumplir la condición.

Una vez que se han calculado las proyecciones sobre el 1$^{er.}$ Componente Principal, se procede a calcular las correspondientes al 2º Componente Principal, para ello se parte de la **matriz de residuales** que se obtiene de restar a la matriz de datos analíticos, ***X*** el producto $\boldsymbol{p.l}^T$.

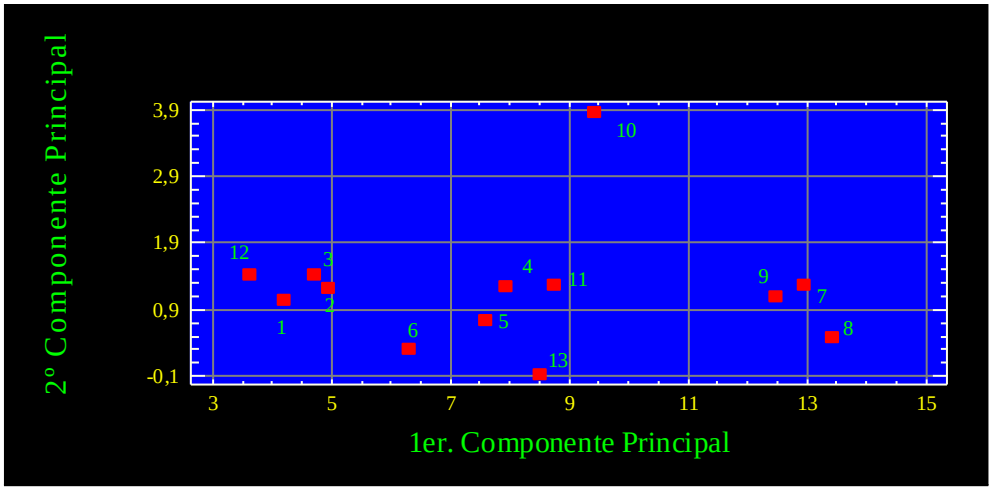
$$E_1 = X - p.l^T$$

El proceso se repite tantas veces como componentes principales se quieran calcular, que será como máximo igual al número de variables o factores incluidos en el estudio multivariante. Si aplicamos este algoritmo NIPALS al ejemplo que estamos utilizando en este estudio, obtendríamos un máximo de tres componentes principales, cuyos "scores" o valores de proyección sobre dichos componentes principales se muestras en la figura.



Como vemos, la mayor información sobre las muestras analizadas se encuentra en el 1er. Componente Principal. Este Componente Principal nos permite ver la forma en que las muestras se relacionan entre sí y los grupos que forman. El 2º Componente Principal nos da una idea aproximada de la dispersión existente entre las muestras, con una (la 10) que se separa bastante del comportamiento general. Finalmente, el 3er. Componente Principal aporta muy poca información, por lo que podríamos prescindir de él y quedarnos solamente con los valores de proyección de los dos primeros componentes principales.

Igual que hemos evaluado estos datos componente por componente, también pueden ser evaluados de forma conjunta representando los valores de las proyecciones sobre el 1er. Componente Principal frente a los valores de las proyecciones sobre el 2º Componente Principal con la que se obtiene una gráfica como la de la figura, denominada **gráfica de dispersión** de las muestras u objetos.



Como vemos, al final del proceso, hemos pasado de un sistema tridimensional de coordenadas a otro bidimensional más fácil de visualizar. Nuevamente las muestras aparecen formando tres grupos igual que en la representación en una sola dimensión. Sin embargo, ahora podemos ver como hay un punto, el correspondiente a la muestra 10, que se separa de las otras muestras sin que podamos asignarlo a ningún grupo en concreto. Esto puede tener dos posibles interpretaciones:

a) Se trata de un dato anómalo producido por algún error durante el proceso analítico (muestra mal tomada, error de medida, etc.) por lo que se trataría de un "outlayer" o valor a no tener en cuenta

b) Puede, por el contrario, tratarse de un único punto perteneciente a otro grupo, del cual solo disponemos de esa muestra.

Igual que se ha hecho en este ejemplo, puede hacerse con series de datos analíticos que incluyan un mayor número de variables a evaluar. Como no es posible su representación gráfica para la posterior evaluación visual, será preciso reducir estos sistemas multidimensionales a sistemas mono-, di-, o tridimensionales, mas fáciles de representar y visualizar. Esta reducción de la dimensionalidad del sistema, se hará como hemos visto, a través de los primeros componentes principales que son los que contienen la mayor parte de la información analítica.

Pretratamiento matemático de los datos analíticos

El ejemplo que venimos utilizando para ilustrar los fundamentos del Análisis Componentes Principales, está formado, como ya hemos visto por 13 muestras u objetos analizados y tres variables, Ca, Fosfato y Ac. Úrico. En estos ejemplos las tres variables tomaban valores de concentraciones del mismo orden (mg/100 mL) y variaban en un intervalo muy similar, de unos pocas unidades hasta 15 aproximadamente, encontrándose, en todos los caso, próximos al origen de coordenadas. Esto, sin embargo, no suele ser la regla, sino más bien la excepción. La tabla de la figura muestra las mediciones realizadas por 10 de las 25 estaciones de control medioambiental del Ayuntamiento de Madrid un día del mes de agosto (26/8/01) para una serie de contaminantes típicos de una atmósfera urbana.

Estación	SO_2, μg/m³	CO, mg/m³	NO_2, μg/m³	Partíc., μg/m3	NO_x, μg/m³	O_3, μg/m³
1	10,6	0,89	88,9	45,2	186,9	32,4
5	13,0	0,69	61,9	62,5	91,2	72,1
6	12,7	1,19	84,9	55,4	217,5	37,4
9	5,8	0,82	95,6	51,8	166,6	28,4
10	15,3	0,77	49,1	33,7	72,5	67,2
12	19,1	0,68	57,4	41,9	94,5	43,4
13	5,0	0,31	46,7	29,8	58,1	63,7
16	6,8	0,42	48,5	38,8	72,9	60,8
19	5,2	0,56	63,7	28,7	95,3	62,1
20	8,4	0,49	52,4	32,8	75,2	57,8

Como vemos en esta tabla, los órdenes de concentración y los intervalos de variación de las concentraciones de los contaminantes son muy dispares. Esto trae como consecuencia que cuando se representan estos datos se produce una importante distorsión de la imagen, siendo, por lo general, muy difícil de apreciar las relaciones existentes entre las diferentes muestras evaluadas. Para evitar este problema lo que se hace normalmente es someter los datos analíticos a un tratamiento matemático previo que nos permita su evaluación posterior en términos relativos.

Existen diferentes procedimientos matemáticos que se utilizan frecuentemente:

Centrado de datos. Es el más simple de todos y consiste en hacer coincidir el origen de coordenadas con el centro del conjunto de datos. Para ello se sustrae a cada variable el valor medio correspondiente a su columna.

$$x^*_{i,k} = x_{i,k} - \bar{x}_k$$

Este pretratamiento es especialmente útil cuando las variables presentan valores en intervalos de variabilidad muy similares pero cuyos órdenes de magnitud son muy diferentes, por

ejemplo, una variable toma valores entre 10 y 20 y otra entre 200 y 210. Sin embargo esto no es muy frecuente, por lo que normalmente el centrado de datos suele ir acompañado de una **normalización** o de un **escalado**.

Normalización. Consiste en hacer que todos los vectores de la matriz de datos tengan la misma longitud, es decir, que la suma de sus cuadrados sea constante. Este proceso elimina el efecto producido por la diferencia en magnitud de los datos originales. El procedimiento consiste en dividir cada elemento de cada vector matricial por la longitud de este vector, la cual viene dada por la raíz cuadrada de la suma de sus cuadrados

$$x^*_{i,k} = \frac{x_{i,k}}{\|x_k\|}$$

siendo $\|x_k\|$ la longitud del vector matricial correspondiente a la columna *k*

$$\|x_k\| = \sqrt{x^2_{1,k} + x^2_{2,k} + ... + x^2_{n,k}}$$

Escalado. Consiste en dividir todas las variables por un factor elegido de tal forma que se compensen las diferencias en el orden de magnitud o en el intervalo de variación de los datos. Las dos opciones más importantes son:

Escalado por intervalo de variación. Consiste en hacer que todos los valores de la matriz queden entre 0 y 1. Esto se consigue restando a cada dato el valor mínimo que toma cada una de las variables (columna) y dividiendo por el intervalo de variación de dicha variable (dif. entre el máximo y el mínimo)

$$x^*_{i,k} = \frac{x_{i,k} - x_k(min.)}{x_k(max.) - x_k(min.)}$$

Este procedimiento de escalado es muy adecuado para detectar la presencia de valores anómalos ("outlayers")

Autoescalado. En este caso, cada variable se transforma mediante centrado de la misma y posterior división por la desviación estándar de la columna en que se encuentra.

$$x^*_{i,k} = \frac{x_{i,k} - \bar{x}_k}{s_k}$$

donde s_k es la desviación estándar de la columna *k*

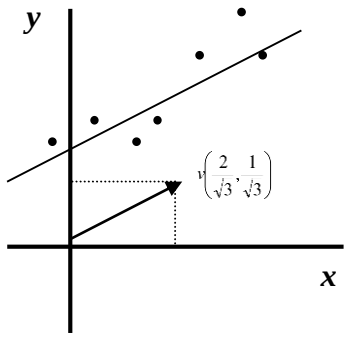
$$s_k = \sqrt{\frac{\sum_1^n (x_{i,k} - \bar{x}_k)^2}{n-1}}$$

De estos procedimientos matemáticos de pretratamiento de los datos analíticos, los más utilizados son el **autoescalado** y la **normalización** con o sin centrado previo de los datos

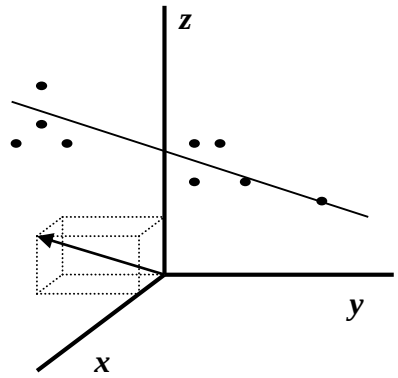
Representación de los vectores de dirección

Como hemos visto, la representación de los valores de proyección y los diagramas de dispersión nos permiten ver como se relacionan entre sí las diferentes muestras analizadas, permitiéndonos apreciar la existencia de similitudes entre las muestras analizadas que hacen que estas se agrupen de una determinada manera.

No obstante, esta información puede mejorarse mediante un cuidadoso estudio de los vectores de dirección. Recordemos que estos vectores van asociados a los valores de proyección y lo que indican es la dirección de cada componente principal en el sistema de coordenadas n-dimensional. Los valores de estos vectores de dirección y su signo nos dan idea de la influencia que tienen unas características sobre otras. .



Así, en el caso más sencillo, el de dos dimensiones en el que se evalúan dos variables *x* e *y* para *n* muestras u objetos si el vector de dirección del 1er. Componente principal fuera p.e. $(2/\sqrt{3}, 1/\sqrt{3})$ esto vendría a decirnos que cuando el contenido del analito *x* o el valor de la característica *x* aumenta, también lo hace *y* pero sólo la mitad.



Igual ocurriría para una serie de objetos analizados si en su evaluación analítica se consideran tres variables o características. Si el vector de dirección fuera $(1/\sqrt{3}, -1/\sqrt{3}, 1/\sqrt{3})$ indicaría que al aumentar el valor de la variable *x*, también lo hace *z* pero *y* disminuye y las tres variables mantienen las mismas proporciones.

Como vemos, los vectores de dirección de los componentes principales nos dan una valiosa información sobre cómo se relacionan entre sí los diferentes analitos o las propiedades incluidas en el análisis multivariante. Por supuesto, esta misma información se obtendría para cualquier otro sistema n-dimensional. Así, para el caso del ejemplo del análisis del aire de Madrid, si mediante el algoritmo NIPALS obtuviéramos las proyecciones de las muestras sobre los dos Primeros Componentes Principales obtendríamos a la vez los dos vectores de dirección que se muestran en el recuadro siguiente:

Row	Component 1	Component 2
1	2,29567	0,584285
2	-0,0278495	-1,46984
3	3,23452	-0,377109
4	2,35308	1,43567
5	-1,10367	-1,38464
6	0,148744	-1,52621
7	-2,43272	0,900288
8	-1,63464	0,378038
9	-1,37623	1,14741
10	-1,45689	0,312097

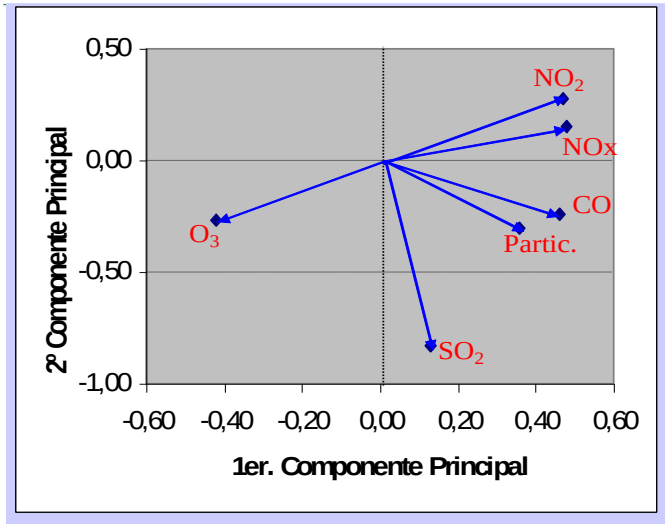
Table of Component Weights

	Component 1	Component 2
SO2	0,131732	-0,813266
CO	0,462641	-0,25427
NO2	0,467806	0,291369
Particulas	0,367622	-0,291428
Oxidos de N	0,4869	0,158482
Ozono	-0,421366	-0,281075

Si observamos los valores que toma el vector de dirección del 1er. Componente Principal vemos que todos son positivos con excepción del correspondiente al ozono. Esto quiere decir que en las muestras de aire analizadas el contenido de ozono disminuye cuando todas los demás contaminantes aumentan y viceversa. Si además observamos atentamente los valores del vector, vemos que todos son del orden de 0.4 - 0.5 mientras que el correspondiente al SO_2 apenas llega a la tercera parte de estos valores. Algo parecido ocurre, aunque no de forma tan acusada, con las partículas en suspensión de <10 μm. Esto quiere decir que en este análisis multivariante, estos dos analitos tienen proporcionalmente menos peso que los otros contaminantes. La conclusión a esta

última observación es simple, la principal fuente de contaminación del aire de Madrid el día en que se tomaron las muestras en las estaciones estudiadas procede del tráfico rodado, ya que la contaminación por SO_2 y la presencia de partículas en suspensión está mas asociada con la combustión de combustibles sólidos, como el carbón, o líquidos, como el gasóleo y el fuelóleo, más utilizados en calderas de calefacción.

Esta información que obtenemos mediante evaluación de los valores numéricos que toman los vectores de dirección también se puede obtener gráficamente mediante representación de estos vectores. Para ello lo que se hace es, sobre un sistema de coordenadas formado por los dos primeros componentes principales, representar como puntos cada uno de los analitos o variables incluidas en estudio, utilizando como coordenadas los valores que para cada uno de ellos toman los vectores de dirección. Después se unen estos puntos con el origen de coordenadas para que tomen el aspecto de un vector. De esta manera se obtiene una representación de los vectores de dirección en un sistema de coordenadas formado por los dos primeros componentes principales.



Un rápido vistazo nos permite confirmar lo que ya habíamos apreciado cuando evaluábamos sus valores numéricos. Todos los vectores se dirigen hacia la derecha (tienen signo positivo) excepto el Ozono que lo hace en sentido inverso.

Otra información que se obtiene de esta representación está relacionada con la pendiente y el módulo de los vectores. Cuanto más horizontales son, mayor es su variabilidad en términos relativos, con respecto al 1$^{er.}$ Componente Principal, es decir mayor es su peso sobre el 1$^{er.}$ Componente Principal, que por otro lado, como ya hemos visto es el que contiene mayor información sobre el sistema analítico en su conjunto. En cuanto al módulo de estos vectores, cuanto mayor es la distancia entre el origen y el punto, mayor es la importancia que tiene esa característica o analito en el conjunto de las muestras. En este ejemplo vemos que los vectores correspondientes al NO_2, NOx, CO y O_3 son los que presentan menor pendiente lo que indica que son los contaminantes que tienen mayor efecto sobre los agrupamientos o similitudes que puedan producirse entre las diferentes muestras, pero por otro lado, el mayor módulo y menor horizontalidad corresponde al SO_2, lo que indica que la mayor o menor dispersión de las muestras depende principalmente de este contaminante.

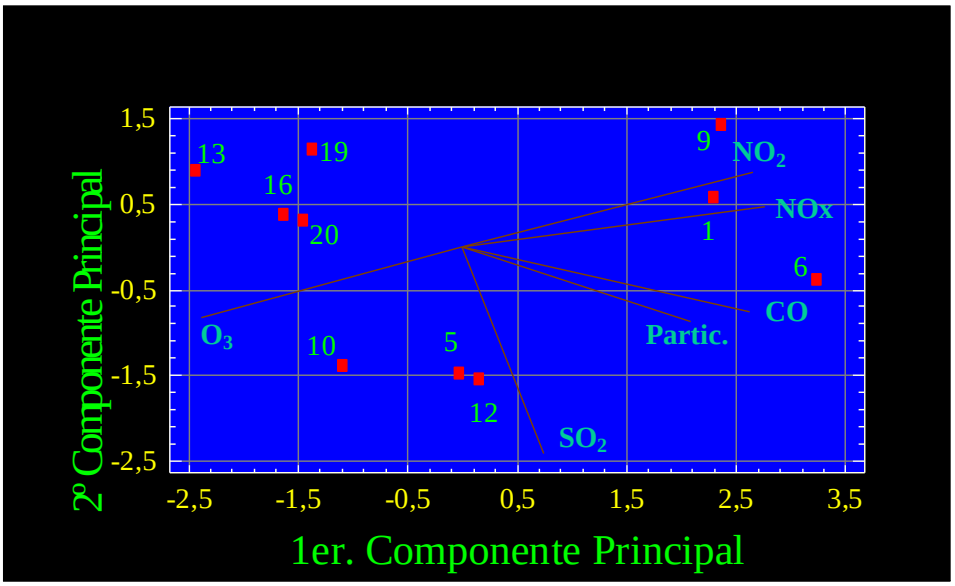
Además de estas conclusiones, que en cierto modo ya las habíamos deducido a partir de los datos numéricos, de las representaciones de vectores de dirección también se pueden sacar importantes conclusiones relacionadas con las correlaciones internas que pueden presentarse entre los diferentes analitos o variables incluidas en el estudio. Vemos en la gráfica que el NO_2 y NOx

presentan vectores muy similares en módulo y entre ambos forman un ángulo muy agudo. Esto indica que estos analitos están fuertemente interrelacionados entre sí, de forma que cuando uno aumenta en una proporción dada el otro lo hace de manera similar. Esto viene a decirnos que tanto el NO_2 como el NOx proceden de las mismas fuentes de contaminación, en este caso, de los motores de combustión interna de los coches, que son la principal fuente de óxidos de nitrógeno de origen antropogénico.

Por el contrario, las direcciones de estos dos polutantes son diametralmente opuestas a la del ozono, esto puede indicarnos, o bien que las causas que hacen que uno aumente hacen que los otros contaminantes disminuyan o también, que existen posibles interacciones entre el ozono por un lado y los oxido de nitrógeno por otro, que es lo que en realidad ocurre, pues el ozono oxida a los óxidos nitrosos pasándolos a óxidos nítricos

Representación combinada de valores de proyección y vectores de dirección

Como vemos, podemos obtener una muy importante información sobre un conjunto de datos a partir de los diagramas de dispersión así como de las representaciones de los vectores de dirección. Como hemos visto, ambas representaciones tienen en común que se realizan tomando como ejes de coordenadas los primeros componentes principales. Esto nos lleva a que, en los casos que sea necesario, puede obtenerse de una sola vez toda la información posible sobre los objetos analizados y las variables incluidas en el análisis mediante representaciones conjuntas de las dispersiones de los objetos analizados y de los vectores de dirección.



La forma que toman estas gráficas combinadas son como las de la figura. Si observamos ésta atentamente, vemos que las muestras 1,9 y 6 se encuentran en la misma dirección que los vectores correspondientes a los óxidos de nitrógeno. Esto quiere decir que los mayores contenidos de estos contaminantes se encuentran en las muestras tomadas en las estaciones 1, 9 y 6. Si observamos la tabla con las cantidades encontradas vemos que, en efecto, esto es lo que ocurre. Por el contrario, al encontrarse estas muestras en la dirección totalmente opuesta al vector del ozono es de esperar, como así ocurre, que estas estaciones son las que menos concentración de ozono han dado. Semejantes conclusiones pueden deducirse para otros contaminantes y otras estaciones.

Como vemos, los gráficos combinados de Componentes Principales permiten una evaluación más rápida de los datos analíticos y suministra información sobre:

a) Relaciones o contrastes entre las muestras u objetos analizados (en este caso diferentes estaciones medioambientales)
b) Relaciones o contrastes entre las variables (en este caso los contaminantes monitorizados en Madrid)
c) Las relaciones existentes entre las variables y los objetos (en este caso los contaminantes y las estaciones donde se tomaron las muestras)

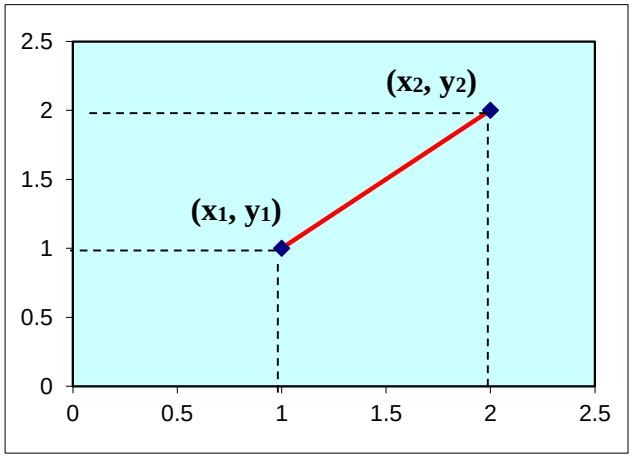
En definitiva y como podemos ver, el **Análisis de Componentes Principales** es una potente herramienta quimiométrica que nos permite visualizar de una forma rápida y sencilla la forma en que las diferentes muestras analizadas se relacionan entre sí, permitiéndonos ver las similitudes o las diferencias existentes entre ellas. También nos permite ver y apreciar las relaciones existentes entre las muestras analizadas y las variables o analitos que caracterizan dichas muestras, permitiéndonos establecer hipótesis sobre las causas que dan lugar a estas interrelaciones.

3.- ANÁLISIS DE AGRUPAMIENTOS

No siempre es necesario realizar un análisis multivariante tan completo y complejo como es el Análisis de Componentes Principales. Cuando el principal interés radica en buscar las analogías y diferencias entre muestras, viendo la forma de agruparse, pero sin importar las relaciones existentes entre las variables, se puede llevar a cabo lo que se denomina un **Análisis de Agrupamientos**.

Puesto que las similitudes y las diferencias entre muestras están en relación con su proximidad o alejamiento en el espacio de coordenadas, lo que hace el Análisis de Agrupamientos es ordenar las muestras en función de las distancias que existen entre ellas. Recordemos que las distancia entre dos puntos en un sistema bidimensional de coordenadas viene dado por la expresión:

$$A\overline{B} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$



que puede generalizarse para cualquier sistema de *n* dimensiones:

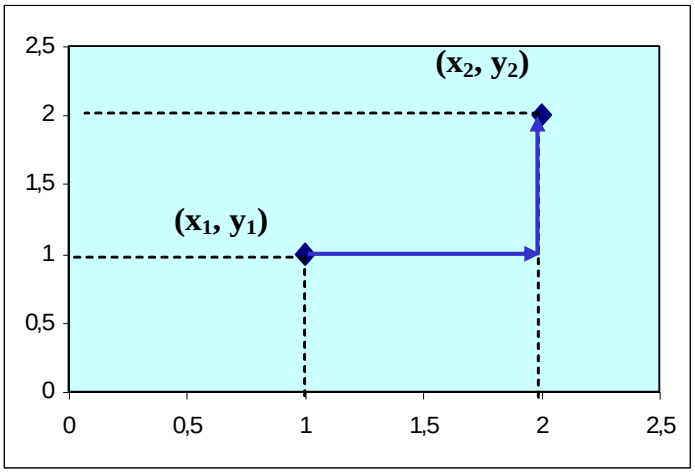
$$A\overline{B} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 + ... + (r_1 - r_2)^2}$$

donde *x*, *y z ... r* representan cada propiedad o variable, medida para cada muestra. Esta forma de medir distancias entre muestras, recibe el nombre de ***distancia de Euclides***. Existen otras formas de medir las distancias entre muestras analizadas, aunque son de uso menos frecuente. Entre ellas destacan la ***distancia de Mikowki***, similar a la anterior, pero que tiene en cuenta el número de variables o factores incluidos en el análisis multivariante:

$$A\overline{B} = \sqrt[p]{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 + ... + (r_1 - r_2)^2}$$

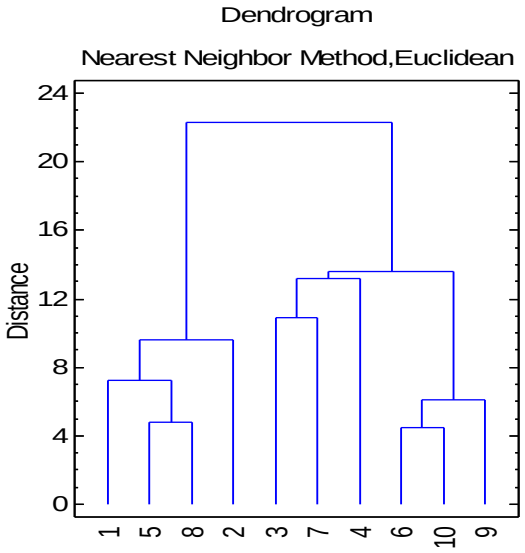
donde *p* es el número de factores o variables y por consiguiente el número de dimensiones. Se aprecia fácilmente que en el caso de que el análisis multivariante se realiza con dos factores o variables, *p* = 2, esta distancia se corresponde con la de Euclides. En otros casos se utiliza la denominada ***distancia Manhattan o City-Block***, más fácil de calcular al no requerir el empleo de raíces.

$$A\overline{B} = |x_1 - x_2| + |y_1 - y_2| + |z_1 - z_2| + ... + |r_1 - r_2|$$



Utilizando cualquiera de estas expresiones se calculan las distancias entre cada muestra y todas las demás y una vez que se han calculado se representan estas en un diagrama jerarquizado que recibe el nombre de **dendrograma**. En la figura de la página siguiente se muestra el dendograma obtenido mediante distancia de Euclides de diez muestras de agua, cuyos valores de dureza, expresados como mg/L de Ca y Mg, se indican en la tabla.

Muestra	Ca (mg/L)	Mg (mg/L)
1	7,4	2,2
2	27,6	6,3
3	55,0	25,4
4	78,0	24,0
5	13,6	6,0
6	50,1	8,2
7	65,1	21,2
8	18,2	4,5
9	55,1	11,8
10	49,0	12,5



Puede apreciarse que las muestras 1, 5 8 y 2 muestran un agrupamiento claramente destacado del resto y que esta últimas se conforman en dos subgrupos bien definidos, muestras 3,7 y 4 por un lado y 6,10 y 9 por el otro.

Como en el caso del Análisis de Componentes Principales, cuando los ordenes de magnitud de las variables incluidas en la evaluación, o sus intervalos de variación son muy dispares, se puede proceder a una normalización u otro tratamiento matemático. También es frecuente en el Análisis de Agrupamientos utilizar las distancias elevadas al cuadrado, lo que permite una mejor visualización de estas, y por consiguiente de los grupos que se forman.

4.- REDES NEURONALES ARTIFICIALES

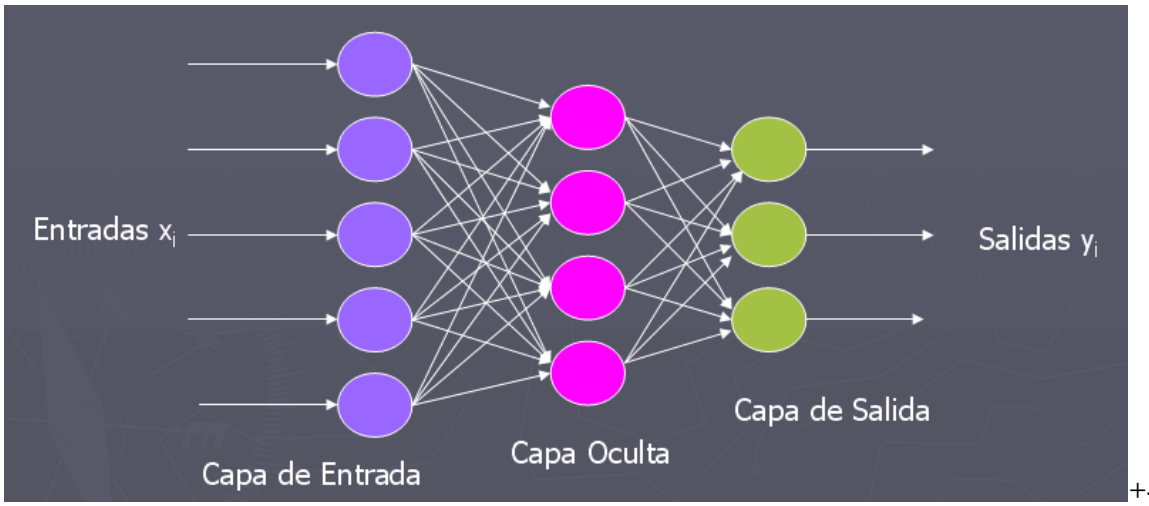
El Análisis de Componentes Principales y el Análisis de Agrupamientos, que acabamos de estudiar, son métodos de análisis multivariante no supervisados, cuya principal función está encaminada a estudiar las similitudes y diferencias que pueden existir entre las diferentes muestras analizadas para poder relacionarlas con algún factor externo que hace que las muestras se agrupen de una manera dada y no de otra forma. Como ya hemos comentado, existen otros métodos de análisis multivariante en los cuales ya se conocen de antemano las causa que hacen que las muestras analizadas se agrupen de una manera u otra. En estos métodos lo que se busca es poder relacionar muestras desconocidas con otras cuyo origen es perfectamente conocido. Son los denominados métodos multivariantes supervisados, también conocidos como **métodos clasificatorios**. De los diferentes métodos clasificatorios que se utilizan, uno de uso muy frecuente es el conocido como Redes Neuronales Artificiales (ANN), que es un método no paramétrico, basado en la estadística bayesiana en el que, además de buscar las analogías entre los objetos, los clasifica según una serie de patrones a grupos previamente establecidos.

Para poder asignar correctamente los diferentes objetos evaluados a los grupos preestablecidos, además de la matriz de datos cuya estructura es la misma que la utilizada en los métodos estudiados anteriormente, tantas filas como objetos evaluados y tantas columnas como variables incluidas en el estudio, se requiere una matriz adicional en el que mediante valores

numéricos se representen las características de cada grupo y deberá tener tantas filas como objetos evaluados y al menos tantas columnas como grupos existan. Por razones de simplicidad en el cálculo matemático, los valores que caracterizan a cada uno de los grupos predefinidos serán binarios (ceros y unos) o polares (unos con signos positivo o negativo). Así, por ejemplo, si todas las muestras a evaluar pertenecen a tres grupos; A, B, y C, a cada uno de estos grupos se le asigna un vector matricial, que debe ser diferente a los otros, como [1, 0, 0] para el grupo A, [0, 1, 0] para B y [0, 0, 1] para C

Matriz de datos					Matriz de grupos		
0.3	1.0	2.0	….	0.8	1	0	0
0.4	1.2	2.5	….	0.7	0	1	0
0.2	0.9	1.9	….	1.0	1	0	0
.	.	.	….	.	.	.	.
.	.	.	….	.	.	.	.
.	.	.	….	.	.	.	.
0.5	1.5	2.8	….	0.7	0	0	1

El nombre de Red Neuronal viene porque, tanto la estructura computacional como la forma en que la red discierne entre unos grupos u otros, se asemeja al funcionamiento del sistema nervioso. La figura muestra un esquema de una Red Neuronal Artificial sencilla, con una capa de entrada, que representa las terminaciones nerviosas de los sentidos, a través de los cuales se adquiere la información, una capa oculta, donde la información es procesada y que a nivel biológico correspondería al cerebro de los animales y por último, una capa de salida, donde la información procesada es trasmitida a otros animales de la misma especie. Como en una red neuronal biológica, estas capas están unidas entre sí por medio de diferentes conexiones, que también reciben el nombre de **sinapsis**.



+-

.

Obviamente, la principal diferencia con las redes neuronales biológicas está en la forma en que la información se almacena y se procesa. Las neuronas artificiales pueden ser de tipo **lógico** o de tipo **continuo**, según el tipo de valores que almacenan:

Lógicas. Solo admiten dos valores posibles

Binaria: $y_1 = [1, 0, 0]$; $y_2 = [0, 1, 0]$; $y_3 = [0, 0, 1]$

Bipolar: $y_1 = [1, -1, 1]$; $y_2 = [1, 1, -1]$; $y_3 = [-1, 1, 1]$

Analógicas o continuas. Admiten cualquier valor dentro de un rango determinado.

El procesado de la información se lleva a cabo mediante funciones matemáticas que permiten trasmitir la información entre las neuronas de las diferentes capas. Estas funciones, a veces llamadas **sinapsis**, son funciones matemáticas de ponderación que establecen la probabilidad de acertar o errar cuando un objeto es clasificado y se estiman en función de los valores de las variables de cada objeto, muchas veces mediante modelos sencillos, como el que se muestra a continuación:

$$Net_j = \sum_{i=1}^{n} x_i w_{ij}$$

El primer paso en una Red Neuronal Artificial, como en las naturales, es el **aprendizaje** o **entrenamiento** de la red. Consiste en procesar una serie de muestras u objetos pertenecientes a grupos previamente definidos y asignar a cada uno de ellos un valor de ponderación que estima la probabilidad de acierto o de error en la asignación de futuras muestras desconocidas. Esta ponderación se calcula en función de los valores de las variables en cada objeto o muestra analizada.

$$w_{ij} = x_i^T y_j$$

donde x_i representa cada uno de los objetos analizados y contiene las características o factores que lo definen e y_i el vector que representa al grupo al que pertenece el objeto.

Después, las ponderaciones obtenidas para cada objeto o muestra, se almacena en una matriz de pesos

$$\boldsymbol{W} = x_1^T y_1 + x_2^T y_2 + \ldots + x_n^T y_m$$

Una vez que la Red Neuronal ha sido debidamente entrenada, es decir, se ha calculado el valor de la matriz $\boldsymbol{W}$, ésta estará en disposición de asignar correctamente una muestra desconocida a alguno de los grupos previamente definidos.

El reconocimiento de la pauta o similitud de un objeto no asignado o desconocido, $\boldsymbol{x}$, se lleva a cabo mediante un proceso iterativo que, en líneas generales, es como sigue:

Paso 1. Se estima un vector de respuesta (grupo al que pertenece).

$$\boldsymbol{y} = \boldsymbol{xW}$$

Paso 2. Se compara el vector $\boldsymbol{y}$ obtenido con los vectores $\boldsymbol{y_1, y_2 \ldots y_m}$ que representan los grupos.
Paso 3. Si $\boldsymbol{y}$ es similar o proporcional a alguno de los $\boldsymbol{y}_i$ vectores que representan a los grupos, se asigna grupo y el proceso acaba. En caso contrario se recalcula un nuevo vector $\boldsymbol{x}^{(1)}$ y se vuelve al paso 1

$$\boldsymbol{x}^{(1)} = \boldsymbol{yW}^T \quad \rightarrow \quad \boldsymbol{y}^{(1)} = \boldsymbol{xW}$$

$$\boldsymbol{x}^{(2)} = \boldsymbol{yW}^T \quad \rightarrow \quad \boldsymbol{y}^{(2)} = \boldsymbol{xW}$$

$$\ldots\ldots\ldots\ldots$$

$$\ldots\ldots\ldots\ldots$$

$$\boldsymbol{x}^{(n)} = \boldsymbol{yW}^T \quad \rightarrow \quad \boldsymbol{y}^{(n)} = \boldsymbol{xW}$$

El proceso se repite iterativamente hasta que aparece la similitud. Si tras un número determinado de iteraciones no es posible asignar grupo, el proceso se detiene asignando la muestra al grupo más parecido o declarándola como desconocida.

Otro aspecto importante en el empleo de las Redes Neuronales Artificiales es determinar la robustez o fiabilidad predictora de la red. Generalmente, como paso previo a ser utilizada para asignar grupos a muestras desconocidas, conviene llevar a cabo pruebas de fiabilidad. Como en el caso de otros métodos multivariantes, estas pruebas se basan en determinar las desviaciones entre el modelo y los resultados experimentales. Un procedimiento habitual, utilizado por muchos de los programas de cálculo estadístico que incluyen ANN, consiste en no utilizar todas las muestras conocidas como muestras de aprendizaje, sino que se reservan algunas de ellas, denominadas muestras de validación para comprobar si existe concordancia entre el grupo al que pertenecen y el que asigna la red. Otro procedimiento menos habitual consiste en eliminar un grupo completo y entrenar la red con los otros grupos. Terminado el aprendizaje se utiliza el grupo eliminado en el proceso de predicción. La red deberá dar como resultado que todas las muestras utilizadas pertenecen a un grupo desconocido.

BIBLIOGRAFÍA

- **LIBROS**

Massart, D.L.; Vandeginste, B.G.M.; Buydens, L.M.C.; de Jong, S.; Lewi, P.J. y Smeyers-Verbeke, J.
Handbook of Chemometrics and Qualimetrics.
Elsevier Amsterdam, 1997.

Otto, M.
Chemometrics. Chemometrics: Statistics and Computer Application in Analytical Chemistry, 3rd ed.
Wiley-VCH: Weinheim, Germany, 2017.

Massart, D.L.; Vandeginste, B.G.M.; Deming, S.N.; Michotte, Y. y Kaufman, L.
Chemometrics: a textbook.
Elsevier Amsterdam, 1988.

Brereton, R.G.
Chemometrics. Data Analysis for the Laboratory and Chemical Plant.
Wiley. England, 2003.

- **ARTÍCULOS**

Pérez-Arribas, L.V., León-González, M. E., Rosales-Conrado, N. **Learning Principal Component Analysis by Using Data from Air Quality Networks**
J. Chem. Educ. 2017, **94**, 458-464

Kryger, L. **Interpretation of analytical chemical information by Pattern Recognition methods. A survey**
Talanta, 1981, **28**, 871-887

- **PROGRAMAS DE ORDENADOR**

ORIGIN Pro 2016 MicroCal Software, Inc.

STATGRAPHICS Centurion XVII Statistical Graphics Corp.

MATLAB R2016a The Mathworks, Inc.